DEVELOPMENT OF INTELLIGENT ALGORITHMS FOR SMART GRID DATA ANALYTICS

A SYNOPSIS

Submitted by

J.JEYARANJANI

(Reg. No. 201507230)

in partial fulfillment for the award of the degree

of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

KALASALINGAM ACADEMY OF RESEARCH AND EDUCATION

(DEEMED TO BE UNIVERSITY)

ANAND NAGAR, KRISHNANKOIL – 626126

TAMILNADU, INDIA

June 2022

1.INTRODUCTION

Smart Grid (SG) is a system of information and communication applications integrated with generation, transmission, distribution, and customer end-use technologies. The base layer of the Smart Grid is the robust electrical network with adequate energy resources. SG enables the shift to more sustainable technologies such as distributed generation and micro grid [1]. Smart Grid system enables the following functions: customer participation to conserve energy, use the current data to address the future sustainability, reduce cost of energy by optimizing the energy flow, integration of communication and control across the energy system [2]. It also overcomes the issues with the conventional system like centralized electricity energy source, uni-directional energy flow, passive consumers participation, real time monitoring.

The smartness is built over the electrical network by the Advanced Metering Infrastructure (AMI) based on Smart Meter (SM) and Wide Area Monitoring System (WAMS) based on Phasor Measurement Unit (PMU). AMI is an integrated system of Smart meters, communication networks and meter data management system that enables two way communications between the utilities and consumers. Smart meters measure and collect electricity consumption data at a frequency of every 15 minutes to one hour. Smart meter is connected to the communication device called the smart meter gateway. Smart meter gateway communicates the electricity consumption information to the next level gateway in the SG as well as to utility. Meter Data management contains the consumer details, real time electricity energy flow measurement data integrated with time of interval[3]. Utility companies may have access to additional data about their consumers socio economic details, e.g., location, square footage of the home or family size, status of the family, chief income amount, etc.. [4]. SM can perform the operations like billing, data management etc., with the help of AMI technology [5]. The deployment of AMI in distribution systems provides a unique opportunity to extend the monitoring capability to the grid edges. It involves many advantages that include consumer's choice to manage their electricity usage and choose the most economically efficient offerings.

Billing is no longer the only function of smart meter. It provides service to all consumers category like resident, commercial and others. The smart meter data can be employed to enhance the efficiency and sustainability of the electric grid. It has diverse characteristics based on the utility company who is implementing the meters [6]. Unprecedented amounts of data collected from the interconnected network are the source of input for meter data analytics. Smart meter data can be used to analyze consumers load curve. The consumer's behavior curve is the overlooked

factor in previous research works. The behaviors vary from consumer to consume, especially for the resident community. The resident's electricity usage hours, life pattern, family composition, and behavior may change over. This perpetuates the change in the load curve of same consumption community. This demographic data can possibly address an electric meter installed premise for investigation. Rather depending on the collection of smart meter data alone, the demographic data can also be utilized in residential energy management.

1.1.Motivation

The key application of residential electricity consumer's data analytics system is Load Prediction, Electricity theft detection, Load Scheduling and Demand Response (DR). A variety of models have been proposed to perform prediction of electricity consumption, consumers load scheduling and consumers consumption pattern analysis and personalized consumer profiling in Demand Response program. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. With the advent of machine learning techniques and evolutionary algorithms, researchers now have powerful tools to further exploit the undiscovered values of residential consumers of electricity. Machine learning algorithms find natural patterns in data that generate insight and help us make better decisions and predictions. An evolutionary algorithm use mechanism inspired by nature and solves problems through processes that emulate the behaviors of living organisms.

1.2.Research Objective

The following are the objectives:

1. To accurately predict the load demand of residential consumers using consumers socio-economic data.
2. To effectively detect the electricity theft using the Smart Meter data.
3. To optimally schedule the electricity load usage of the residents who participate in Demand Response program.

2. LITERATURE SURVEY

The modernization of electrical grid has the opportunity to optimize the problem solutions of electricity utility companies and the consumers. In the area of Smart Grid, researchers adopt several techniques and methodologies to address the existing issues that improvise the working model. The SG Data analytics work includes load profiling, load prediction, theft detection etc.

To classify the attack types and causes for the non-technical loss (NTL) in AMI, features and feature-engineering methods [7]. A combined model of customer energy consumption pattern classification method based on probabilistic neural network and mathematical model based on Levenberg- Marquardt method to detect the non-technical loss is proposed in [8] .Review of methods based on algorithms implemented, features extracted, and metrics used for evaluation for NTL detection in recent literature [9]. Also based on the time of response, performance result, it presents the comparison of categories of data-based, network-based, and hybrid methods. An ensemble deep learning network based approach to detect the illegal electricity consumption of the consumers using AdaBoost and Deep Neural Network method in continuous time series data [10]. Their experimental result exposes that the ensemble model outperforms the Support vector Machine, Artificial Neural Network and Random Forest models. A model to identify the malicious consumers using feature engineering to extract new feature from raw dataset, clustering to find the subset of feature that influence the electricity consumption, and classification methods such as decision tree, random forest, and gradient boosting to identify the consumer is genuine or malicious [11]. The tree based classification algorithms performances are compared and found that gradient boosting results better. Electricity theft implemented by balancing mechanism based on the interquartile minority oversampling technique and a combined model (long short-term memory, UNet and adaptive boosting) to overcome the shortcomings of conventional approaches [12]. The performance indicated using area under the curve, precision, recall and F1 measure shows that it works well for theft detection.

The synthesis method to characterize the consumption patterns within a daily domestic demand pattern, informed estimations can be with regards to the type of activity being carried out within the dwelling of specific country [13]. The result provides investigation on comparison of the synthesized aggregated datasets with real demand profiles from substations.

To perform load scheduling of residential consumers by firstly classified into different categories according to various Demand Response programs [14]. Then scheduling is modeled based on load supply and demand of resident. The model result presents reduction in electricity consumption cost of the residential consumers. Through this DR program, the residents participate and decrease their peak-valley load difference value efficiently. To overcome the future uncertainties in electricity generation using incentive-based DR with reinforcement learning and Deep Neural Network to optimally predict the incentive rate for every specific consumer [15]. From the result it is inferred that the rate of participation of consumers in DR

program is increased. A case study using real time data from Egypt to compare the performance of mixed integer linear programming model with cuckoo model and proved that cuckoo model provide cost saving that range from 57% to 80% [16]. To provide load scheduling for residents and developed a Home Energy Management System to address the energy demand problem of utility companies and the real time cost problem of the consumers through DR program strategies. The investigation result shows that there is reduction in time granularity [17].

The optimization of electricity consumption cost of the consumers for DR program. The learning model is built to identify the different consumption pattern of the appliances in resident [18]. Genetic algorithm based distributed pricing scheme is proposed to maximize the profit of the retailers. Their numerical result justifies the model performance. A model to optimize the load of Home Energy Management System [19][20]. Artificial Neural Network and Genetic algorithm together is modeled to find the best combination of loads in resident to reduce the energy demand period. The solution is stress-tested and demonstrated for four bedroom house with grid energy usage reduction by 10%, 25%, and 40%.

From the previous research works on Smart Grid, it is observed that Load prediction, electricity theft detection, Load Scheduling and Demand Response are the notable problem in Residential Smart Meter Data Analytics System (SMDAS). Real time SM data along with their socio-economic data is found to contribute in designing an effective model. It is essential to provide a model to address these issues and improvise the performance. Machine Learning and evolutionary algorithms are efficient methodology to implement and evaluate the model performance in terms of accuracy and error [21][22]. The model performs classification, clustering, prediction and optimization. The residential consumer's electricity consumption pattern varies based on their socio-economic, season, environmental factors and also the real time electricity cost that change on time. This increases the risk of electricity energy demand that increases the complexity of the smart grid with unplanned electricity generation. This problem leads to build a sustainable Residential AMI based on Smart Meter Data Analytics System (SMDAS) model for load prediction, identify the non-technical loss, schedule the loads of the resident, and reduce the electricity energy demand. Upon concentrating on this, the research objective is formulated.

3. RESEARCH PROBLEM

3.1. Dataset

A smart Grid project initiated with the automation, monitoring, and two-way communications technology to improve the electric grid function in Ireland in 2009. The Commission for Energy Regulation publicly released full data sets of "Electricity Smart Metering Customer Behavior Trials" in Ireland recorded from July 14, 2009 to December 31, 2010 and with over 5000 Irish homes and businesses. We obtain the dataset through Irish Social Science Data Archive (ISSDA) on 2017. The SM data's are available for both the residential and commercial consumers. The SM devices are installed on the resident's home records the electricity consumption data for every 15 minutes. The socio-economic data also called as demographic data of those residents are collected as survey questionnaires. Data preprocessing techniques are applied to transfer the raw data into understandable format.

3.2. Load Prediction

3.2.1. Estimating residential building energy consumption using demographic data

Load forecasting is the prediction of the future load demand of electricity consumers. Power distribution companies rely on short and long-term forecasts at the feeder level to support operations and planning processes. Smart meters make it possible for the distribution companies to better understand and forecast the load of an individual resident or building. The demographic data of resident's are collected through 235 survey questionnaire and their electricity consumption data collected from smart meter. The fine-grained smart meter data introduces new perspectives to the aggregation level load forecasting. Data preprocessing like scaling, handling missing data, encoding categorical data are performed.

Initially, K-Means clustering which is an unsupervised learning algorithm categorized under partitional clustering technique is implemented. The consumers are grouped based on the similarity of their demographic data. Also it assists in validating the prediction system and initiates the step towards residential electricity detection. Elbow curve method is used to determine the number of clusters (i.e 'K' value) for partitional clustering method. An individual model is developed for every cluster to perform load prediction.

The Neural Network (NN) model is trained with demographic data as input and the electricity consumption as the output. The neural network is built with five layers namely input layer, 3 hidden layers and output layer. The parameters of the network are as follows: Batch size -

400 per epoch, Verbose -1, Activation function - relu, Kernel_initializer - uniform, Optimizer - Adam optimizer, Epochs - 20. The mean square error (MSE) is the loss function used for the prediction performance evaluation. The developed model predicts future electricity consumption value with the Neural Network model of the resident using demographic data rather on smart meter data alone. So this model is generic and can be implemented for the electric meter environment as well where there is no time series data representation. The testing data (new consumer demographic data) is compared with all k-cluster patterns and the appropriated trained model of that cluster is allowed to perform load prediction and to calculate the possible deviation between the actual and the predicted. The Figure 1 and 2 shows the models training and testing phase.

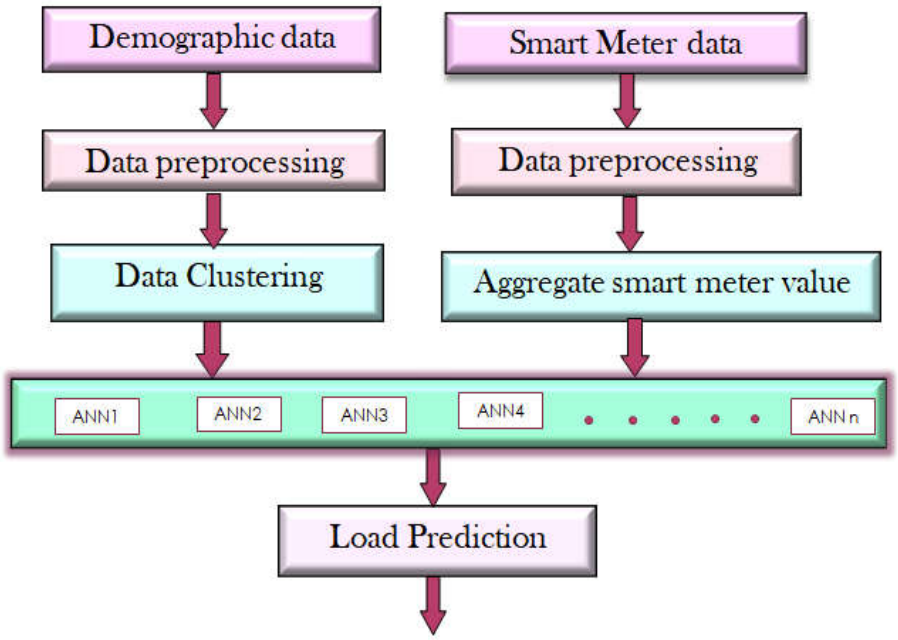

Figure 1. Model Training phase

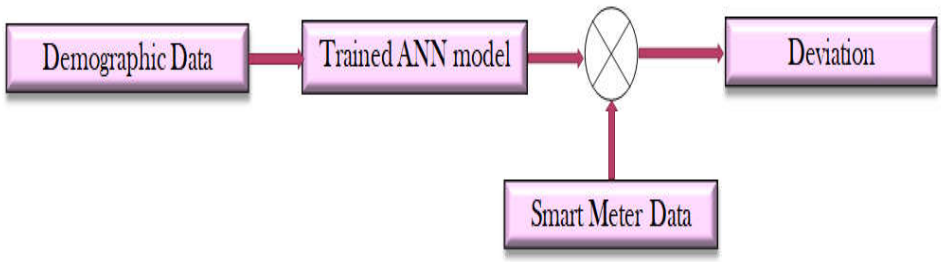

Figure 2. Testing Phase

Table 1 shows the Network layers type, shape and the parameters involved for building the model.

Table 1. NN layer parameters

Layer (type)	Output Shape	Param #	dense_1
(Dense)	(None, 10000)	2370000	dense_2
(Dense)	(None, 1000)	10001000	dense_3
(Dense)	(None, 1000)	1001000	dense_4
(Dense)	(None, 1000)	1001000	dense_5
(Dense)	(None, 1)	1001	

Figure 3 shows the prediction error rate calculated in each epoch of a Neural Network model.

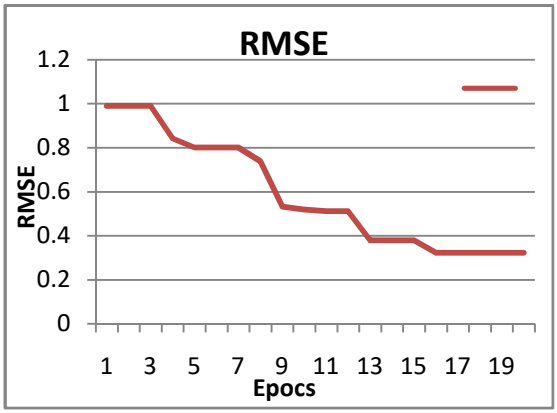

Figure 3. Testing phase RMSE value in each epoc

3.3. Theft Detection

3.3.1. Machine Learning Algorithm for Efficient Electricity Theft Detection

Electricity theft is the most common issue in smart metering environment. The Technical loss is the power dissipation in transmission lines and distribution. Non Technical Loss (NTL) is concerned with unidentified, misallocated or inaccurate energy flow. It is possible to compute the technical loss but not the NTL. The NTL is caused by Meter fault, Incorrect billing, Electricity Theft etc. The theft include bypassing the electricity meter, tampering the meter reading, or hacking the meter. The electricity theft has its impact on the electricity load estimation. The utility companies perform the load estimation with the available load value provided by the Smart Meter. Apart from financial loss, the irrelevant estimation value may lead the utility to undergo imbalance in generated energy and required energy. Therefore, an Efficient Electricity Theft Detection (ETD) module is required to solve the NTL issue of Transmission and Distribution units. Figure 4 shows the overall architecture diagram of the approach.

Firstly, a random 50 numbers of consumers demographic data are selected and considered as genuine profiles. The trustworthiness of the consumers is verified using an k-means unsupervised learning algorithm. As a myth, consumer's electricity consumption will be approximately same who have similar pattern of life style. Then the average of 28 days SM data values of consumer is calculated. The k-means algorithm groups the consumers into 'k' number of clusters. Figure 5 shows the cluster result. The average SM data of all the consumers in a cluster are compared and the result infers that all the consumers average SM data with the difference + 0.5 or -0.5 to that of cluster average. This is the similar case for all the clusters of the experiment. This proves that the considered consumer's profiles are genuine. Secondly, the bogus consumer's profiles need to be generated. Three types of synthetic SM data values namely T_1, T_2 and T_3 are generated for the actual average SM data (D_t) of those 50 consumers profile. 'α' is the random value generated in the range of -0.5, 0.5), 'β' is the random value generated in the range of (0.1, 0.8) and 'γ' is the mean value of the (D_t). these 'α', 'β' and 'γ' are multiplied with the (D_t) to generate the bogus profiles.

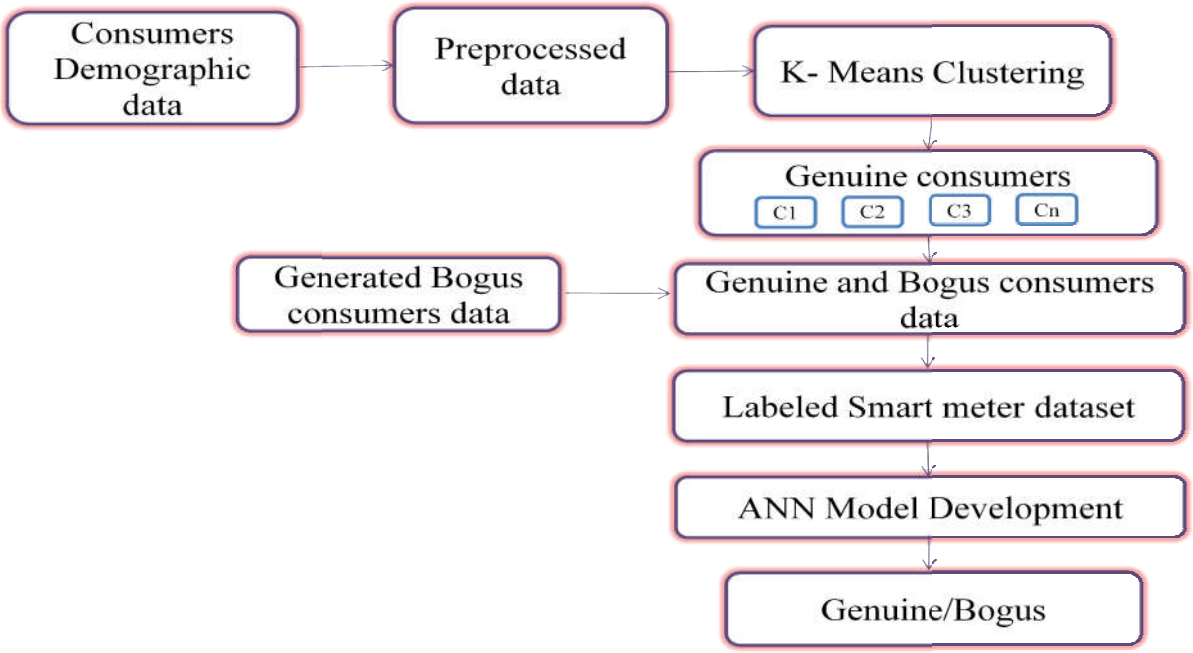

Figure 4. Machine Learning based ETD architecture diagram

Next, the Feed Forward Neural Network based classification model is built to perform the prediction of class. The labeled dataset is prepared by combining the genuine profiles and the generated bogus profiles. The 28 days SM data of the consumers are the input. The output value of the genuine profile is '0', type1 bogus is '1', type 2 bogus is '2' and type 3 bogus is '3". For the network training 80% of the dataset is used and 20% of the dataset is used for testing. The

network uses the weights and bias value for building the classification model and the layers included are input, one hidden and output layers. The neural network model parameter values are as follows: Layers – 3, Batch size – 10, Verbose – 1, Activation function – ReLu, Kernel initializer – uniform, Optimizer – Adam optimizer, Epochs –10. Figure 6 shows the network structure of the model. With the predicted class label value, the model is evaluated using the performance evaluation metrics. The difference in actual class label and the predicted class label are considered for the performance evaluation. The performance of the proposed system uses Accuracy, Precision, Sensitivity and Specificity as performance indices. Table 2 represents the performance evaluation results.

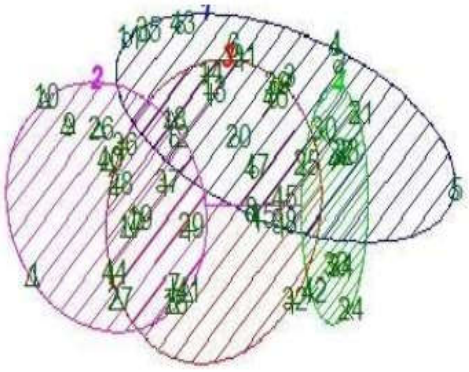

Figure 5. K – Means Cluster

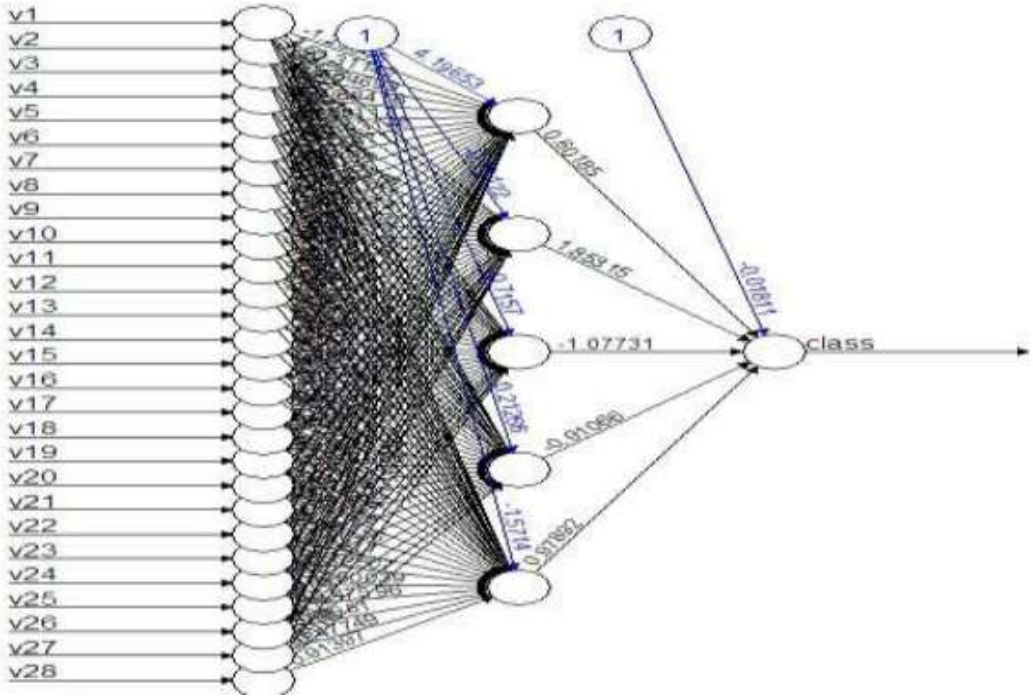

Figure 6. NN Classification

Table 2. Theft Detection performance Evaluation

Class	Precision	Sensitivity	Specificity
0 Actual	0.85	0.85	0.96
1 (Type 1)	0.87	0.87	0.95
2 (Type 2)	0.81	1	0.92
3 (Type 3)	1	0.90	1

3.3.2. An Ensemble Machine Learning Model for Electricity Theft Detection

The improvisation in electric meter to smart meter increases the grid issues in terms of energy theft. So the SM requires an additional layer called cyber layer to address NTL such as meter hacking, tampering and bypassing type of fraudulent activities. There is the possibility of cyber crime during the data transmission, before data being communicated from and to the SM. This work focuses on using residential demographic data of the SM consumers rather than depending on their SM data alone. There are different machine learning approach depending on SM data is found to be inefficient prediction where the energy consumption changes over their social life style. The proposed model is built as five modules. Figure 7 shows the overall architecture of the proposed Ensemble Machine Learning Model for ETD. The system has the following modules: i) Data Preprocessing and Generation ii) Feature Selection iii) Learning Model and iv) Ensemble Learning.

Firstly, real time smart meter data obtained from the ISSDA are considered as genuine. As like the previous work, the bogus SM data's are generated and are combined with the genuine data's. Equation (1), (2) and (3) are used for generating the three varieties of bogus data. The dataset is incomplete and also encounter the missing data values. Preprocessing techniques are used to prepare the dataset suitable for learning. The feature values are in different range values. To scale the feature input, Min-Max scaling technique is implemented. As there are 235 demographic feature input values, the complexity of the learning algorithm may increase. Feature selection technique is the powerful method to reduce the learning complexity and also not missing any useful data. The correlation coefficient based feature selection procedure is carried out to indentify the predominant feature input that involve in electricity energy usage of the resident consumers.

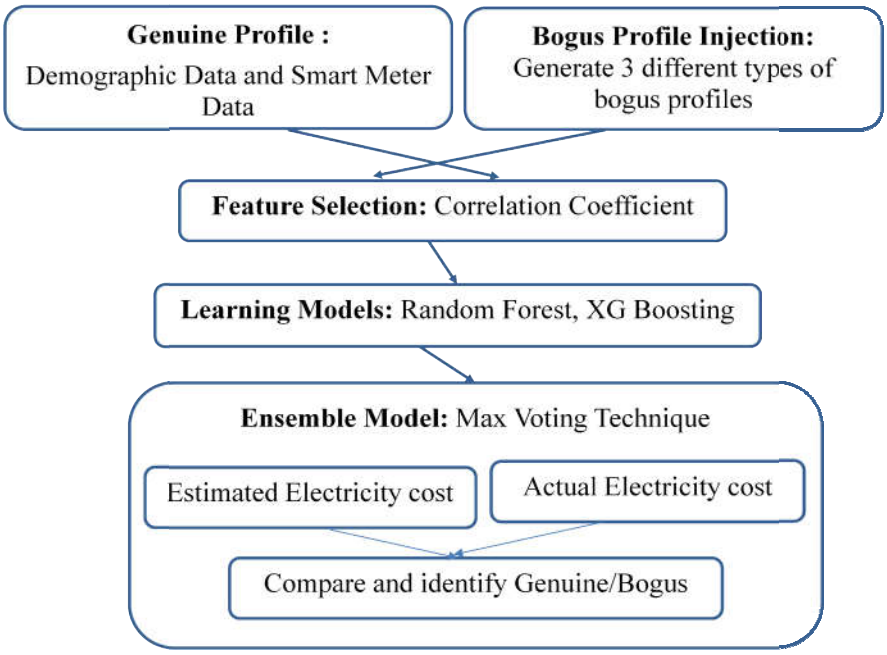
Genuine Profile :
Demographic Data and Smart Meter Data
Bogus Profile Injection:
Generate 3 different types of bogus profiles
Feature Selection: Correlation Coefficient
Learning Models: Random Forest, XG Boosting
Ensemble Model: Max Voting Technique
Estimated Electricity cost
Actual Electricity cost
Compare and identify Genuine/Bogus

Figure 7. Ensemble based ETD Architecture Diagram

$$T1 = C_x + \text{rand} \quad (1)$$

$$T2 = \text{REPLACE}\,(C_x)\ \text{with}\ '0' \quad (2)$$

$$T3 = C_x + \text{mean}\,(C_0, C_1, C_2, C_3, C_4, \ldots\ldots\ldots C_n) \quad (3)$$

Tree based learning models namely Random Forest regression model and Extreme Gradient boosting learning model are used to construct the decision tree for the performing prediction. In random forest regression learning, an ensemble of decision tree which constructs many decision tree in random. It estimates the error rate more accurately compared to that of decision tree. The prediction result is obtained by averaging each of tree's individual prediction. Extreme Gradient boosting learning model is tree based and improved gradient boosting learning model. It regularizes the learning objective to obtain minimum of the loss function. It is faster than the traditional ensemble learning to predict. An ensemble machine learning model is developed to detect the consumer with unusual power consumption pattern. A tree based ensemble machine learning model is built as the combination of Random Forest learning model and Xgboost model. Max voting is a technique where each base model makes a prediction and votes for each sample. Only the sample class with the highest votes is included in the final predictive class.

Two purpose of regression analysis is prediction and to disclose the feature input and output relationship. The different regression models are built using conventional learning models which includes Multi - Linear Regression Model, Stepwise LR Model, Bayesian LR Model, K-Nearest Neighbor Model, Sparse Partial Least Square Model, Elastic-Net Model and Elastic-Net Regularized Generalized Linear Model. The models are trained and tested for it performance and the performance metrics used is Root Mean Square Error. Figure 8 shows the testing performance box plot of the entire considered conventional regression model. This result is used for performing comparative analysis with the tree based ensemble model result that is built in the next module. Table 3 shows the Performance Measure of conventional models.

Table 3. Performance Measure of conventional models

Model	alpha	lambda	RMSE	Rsquared	MAE	RMSESD	Rsquared SD	MAESD
Multi - Linear Regression Model	0.55	0.0644721	1.629054	0.05425647	1.217256	0.0563929	0.02369820	0.03527280
Stepwise LR Model	0.10	0.0644721	1.629058	0.05344718	1.216156	0.0575570	0.02359047	0.03611849
Bayesian LR Model	1.00	0.0064472	1.629148	0.05337824	1.215915	0.0577351	0.02354164	0.03625166
K-Nearest Neighbor Model	0.55	0. 0064472	1.629426	0.05310456	1.216011	0.0579109	0.02350067	0.03639964
Elastic-Net Model	1.00	0.0006447	1.629752	0.05279808	1.216143	0.0580861	0.02342216	0.03655458
Lasso	0.55	0.0006447	1.629754	0.05279603	0.058086	1.216147	0.02342313	0.03655479
Elastic-Net RG Linear Model	0.10	0.0064472	1.629764	0.05277745	1.216174	0.0580850	0.02342080	0.03655550

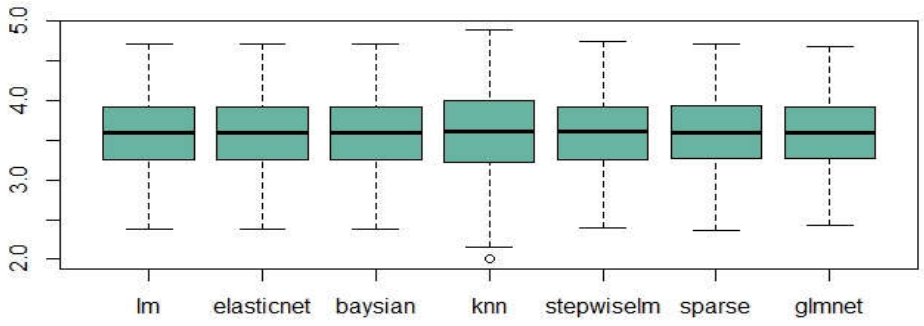

Figure 8. Testing Phase RMSE box plot

Max voting technique based ensemble model is formed by combining two tree based regression model – Random Forest and Extreme Gradient Boosting learning models to perform the prediction of bogus consumers. During the training phase, both the model uses different split of data samples from the same training dataset. During the testing phase, predictions (votes) are provided for each of the testing sample. In max voting, the final prediction output is based on the maximum vote value. The learning summary of the ensemble models is shown in the Table 4. The accuracy value of 99.06% is obtained as the performance metric value of the tree based ensemble model. The mean accuracy value of the Random Forest and Extreme Gradient Boosting are comparatively lesser than the ensemble (RF+XGB). Hence the ensemble model is found to be the outperforming model than the conventional regression model.

Table 4. Ensemble Learning Algorithm evaluation

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
RF +XGB	0.826	0.8562	0.880	0.8956	0.9234	0.9739

3.4. Demand Response

3.4.1. Load Profile Generation

The widespread issue that the electricity utility companies endure is the scarcity of electricity that changes over time. The raise in demand of electricity is predicted in advance to manage the electricity supply accordingly. The predicted demand periods is confessed to the consumers and are expected to reduce electricity during that demand period. The electricity cost is dynamic that changes its rate based on the supply demand. The demand response program is executed by the utility aim to reduce the electricity consumption in demand period. Therefore, the residential consumers whose consumption changes over time are focused for the electric energy related issues like Demand Response. The electricity consumption pattern differs from resident to resident based on number of factors like their location, socio-economic details, environmental factors etc. To investigate certain important factors, a mathematical model for generating the residential consumers load profile is proposed in this paper. It aims to categorize consumers based on dwelling type. Mathematical model is used to generate resident load profile for each dwelling type. The dwelling types of the residents are the building types like bungalow, apartment, Villa, individuals and condominium.

$$P_{start}(A,t) = \sum P_{sat}(A)* N_A * P(A, c) * P_{step} + T_{comp} \quad (4)$$

The model formula represents the resident's appliances and its possible utilization time for generating the load profile specific to each dwelling type. The load profiles are generated based on resident activity. For each activity, certain electric appliance is used by the resident, whose electricity consumption is linked to its corresponding timeslot of the day to generate load profile of the resident. A number 0 or 1 is generated as result, $P_{sat}(A)$ as represented in equation (4). Based on the number of similar appliances (N_A), the number is multiplied. The probability of an appliance on/off state is P (A, c).The probability time slot at which the appliance may be turned on is P (A, t).

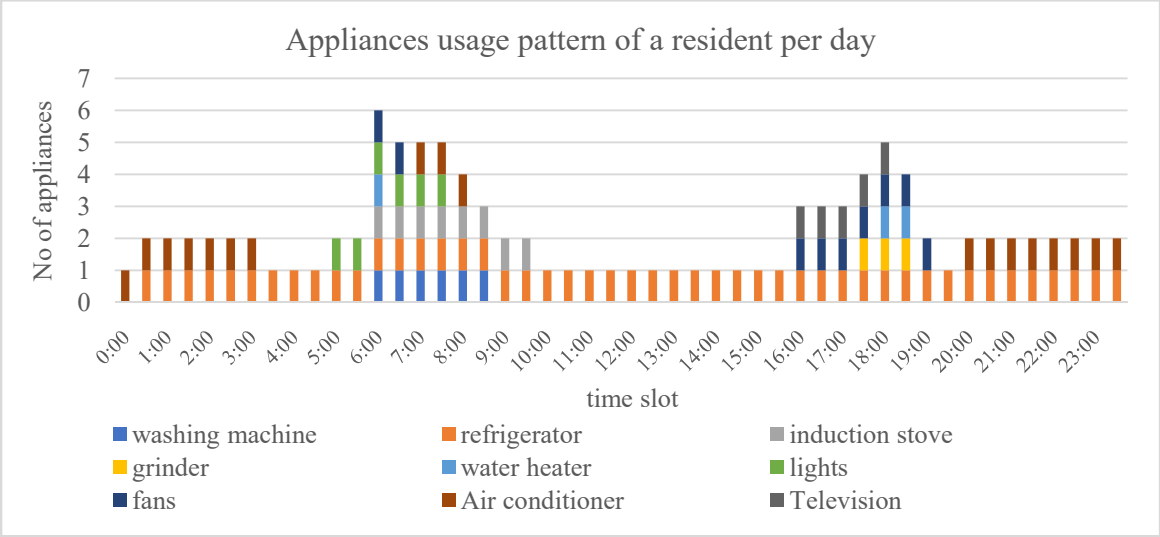

Figure 9. Sample load profile generated for apartment resident based on appliances running status

Figure 9 represents resident's appliances whose operation running status is in 'ON' for a day. It shows the sample load profile generated for apartment dwelling type. The appliances provided for each of the dwelling type are assumed to be available and are in working state. The prediction of average electricity consumption of the resident who participates in DR program is possible. Similarly, this prediction value obtained for all the selected profiles are summed to know the overall demand of future electricity demand. This prediction is used to reduce or vanish the demand period. The demand of electricity is calculated by subtracting the required from generated electricity. The optimal consumers are identified to participate in the DR program of the utility company. The optimal consumers are selected based on their earlier participation in DR program and also based on their planned electricity consumption that is more during the demand period. This reduces the complexity of the DR program by targeting only the possible consumers rather all.

Table 5. Selected Consumers for DR participation

RESIDENT TYPE	SELECTED LOAD PROFILE
Apartment	"A 2" "A 9" "A 11" " A 14" " A 15" " A 17" " A 23" " A 29" " A 37" " A 42" " A 43" " A 50" " A 52" " A 58"
condominium	"S 4" " S 5" " S 6" " S 13" " S 16" " S 18" " S 22" " S 26" " S 27" " S 28" " S 29" " S 31" " S 40" " S 42" " S 43" " S 44" " S 48" " S 59"
Individual	"D2" " D 5" " D 9" " D 10" " D 12" " D 14" " D 16" " D 31" " D 32" " D 38" " D 41" " D 43" " D 48" " D 49" " D 50" " D 51" " D 55" " D 56"
Villa	"T 6" " T 11" " T 13" " T 18" " T 45" " T 47" " T 49" " T 55" " T 56" " T 58" " T 60"
Bungalow	"B 4" " B 8" " B 11" " B 15" " B 16" " B 17" " B 19" " B 20" " B 23" " B 25" " B 28" " B 29" " B 32" " B 35" " B 38" " B 41" " B 44" " B 46" " B 49" " B 51" " B 52"

Table 5 represents the consumers profile selected for participating in DR program. The demand period has a demand start timeslot and end timeslot. The profiles are selected that satisfies the optimal customer selection condition for the DR program. The consumers who are willing to reschedule their appliances operation time and are active participants of earlier DR program. For the considered experiment, the total load generated is 10500 kwh a random demand period as 16:00 to 18:00 for a day.

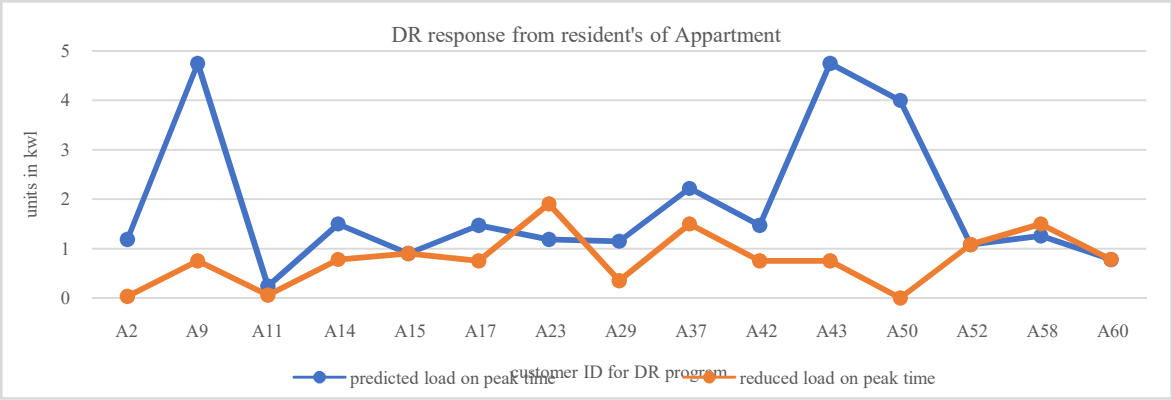

Figure 10. DR response from optimal customer of apartment resident

Figure 10 represents the reduction of consumption by the selected consumers after who participate in DR program. The response received from the residents of the selected consumers profiles by setting the appliances status off or idle for the DR period is the major contribution to the reduction in peak load. It varies from resident to resident, dwelling type to dwelling type. The

synthetic profiles generated through the mathematical model are experimental input for the DR program. These profiles contribute to valley the demand period.

3.4.2. Genetic Algorithm based Resident Load Scheduling for Electricity Cost Reduction

The smart grid is an environment where the electricity utility companies generate electricity energy and transmit it through bi-directional transmission lines to the consumers. There is an unplanned demand or valley in electricity energy usage of the consumers. Load forecasting is the preferable mechanism to avoid these types of demand and valley for efficient energy utilization. The commercial energy requirement are preferably cannot be disturbed for the reason of interrupt in the productivity. The residential consumers are considered whose load values can be altered based demand time. DR is the program run by utility to reduce the demand time of electricity energy. A day ahead energy prediction scheme predicts the requirement of energy from the historic energy consumption. The demand side management proposes the real time pricing schemes that varies the electricity cost based on demand and valley time. If there is surplus energy available in specific time, then the cost per unit is reduced in that time whereas during the demand time, the cost per unit is increased.

The device available in the resident is categorized into two namely uninterruptible and shiftable. To address the demand time, scheduling of residential appliances are preferred by utility and the residents are expected to participate in the DR program. The participation of the residents in DR will reduce their electricity cost and also reduces the energy demand of utility. Penalty is the cost imposed on the resident for using electricity energy on demand time. The timestamp is considered as 24 timeslots per day. The uninterruptible device is provided with the priority for utilizing the energy whereas the threshold energy value of the resident may not be a factor of consideration here. The threshold energy of the resident is verified for scheduling the remaining device such as shiftable devices. This type of scheduling operation may reduce the penalty of the resident. The numbers of considered devices, N = 20, considering that the resident may have more than one same type of device. For our experiment, we considered that also as individual device eg: if a resident has two induction stove, then it will be listed as two loads. The reason behind is the two load working timeslot may be varying. The ‘N’ devices include both uninterruptible and shiftable devices. he main aim is to reduce the electricity const without reducing the usage of electricity energy. The problem is formulated as minimization problem. Population based stochastic optimization approach: Genetic Algorithm is implemented. In real coded GA, the solution values are represented in its natural form. The individual solution has 20 chromosomes

(number of appliances) with 24 genes (number of timeslots). Initial random population is generated and the fitness of each individual is evaluated. The tournament selection method is used for selecting the fittest chromosome for the next generation. Genetic operator's crossover and mutation are employed to fine-tune the results of selection operation. Blended crossover method is applied for the natural form representation of solution. As there are 24 timeslots, one random number is generated between 0 and 24 and is used for mutation. The best individuals are selected in every iteration and the optimal solution is obtained after number of generation. To the minimization problem, 4 numbers of constraints are formed.

1) Threshold energy monitoring is represented in equation (5): For every day, the electricity energy consumption of the resident is verified to be less or equal to the average value.

$$\text{Threshold Energy, Thres\_E} = \{ \sum_{t=1}^{24} \{ \sum_{n=1}^{20} [LSt_n * LEt_n]/20 \} / 24 \quad (5)$$

2) Uncompromised Energy is represented in equation (6): For the purpose of participating in the DR program to reduce the electricity consumption cost, the planned electricity usages of the residents are not compromised.

$$\text{Total required Energy, Tot\_E} = \sum_{t=1}^{24} \sum_{n=1}^{20} [LSt_n * LEt_n] \quad (6)$$

3) Shiftable load on peak period is represented in equation (7): The planned working times of the appliances are altered corresponding to the demand period time.

$$OSLntmax < Pmin \;||\; OSLntmin > Pmax \quad \text{for all } n \in N\_shiftable \quad (7)$$

4) Uninterruptible on peak period is represented in equation (8): The determined electricity energy requirement are proposed by the uninterruptible appliances. The allowed threshold energy levels on every timeslot are utilized by these uninterruptible devices and are not altered in their planned operation time.

$$Tot\_ULnt = \sum_{n=1}^{N} [SULt_n * EULt_n] \quad (8)$$

The objective of the work is to minimize the overall electricity cost of the resident per day and also to schedule the loads according to the dynamic cost as proposed by the electricity utility

company. The appliances are scheduled where the average electricity usage of the resident is not compromised and also the uninterruptable loads are preferred to occupy the timeslots during peak demand. On violating any of the proposed constraints, penalty cost is added to the proposed cost for the dynamic period timeslot.

The penalty is calculated using the equation represented as $P = \sqrt{A - Thres\_E} \quad \forall t$. Where A is the sum of all units of scheduled appliances at timeslot 't'. As this experiment has 24 timeslot per day, 24 penalty values are calculated and added to the dynamic cost of every timeslot. The genetic algorithm is implemented to find the optimal solution that produces less/no penalty. After number of iterations the penalty become neutral or zero. After obtaining this expected output, the generations are stopped.

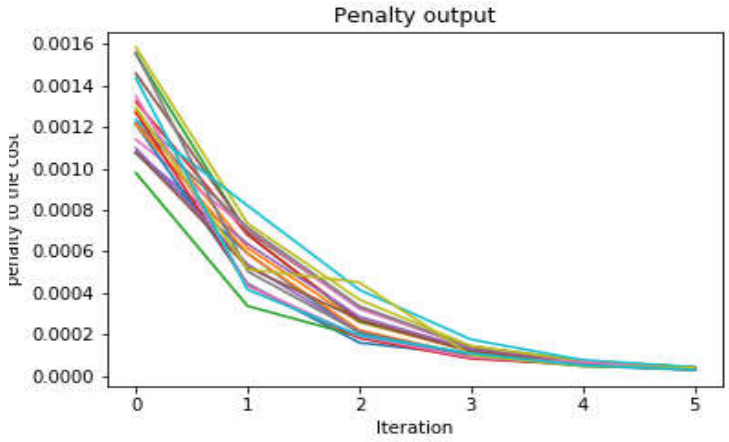

Figure 11. Penalty value generated in each generation of GA

Figure 11 shows that penalty value that is generated during each iteration of the optimization algorithm. Then it is added to the dynamic electricity cost. The genetic algorithm is applied to identify the best individuals in each of the iteration. It is inferred that the penalty value is getting reduced after each generation and by generation 5, the penalty is '0'. This work addresses the demand response program by scheduling the loads of the residents in an efficient way. This reduces the demand of electricity and also the dynamic cost that rises more during the demand period.

3.4.3. Improved Genetic Algorithm for Optimal Demand Response in Smart Grid

In smart grid, variable electricity pricing scheme is adopted. This variation is based on the electricity production rate. Even if the aggregated power consumption of the consumers is same, the Real Time Pricing scheme will generate different price based on the electricity units consumed with respect to the timeslot of the day. The electricity utility company varies the

electricity tariff for every timeslot of day. The electricity price in each hour is represented as C_t. In this work we consider a day as 96 timeslots (15 minutes is equivalent for 1 timeslot).The home appliances considered include washing machine, fridge, fan, laptop, kettle, light etc... The appliance whose operation time is possible to shift its working slot is categorized as the interruptible appliance and whose operation schedule cannot be altered is categorized as uninterruptible appliance. The interruptible appliances are those can have split in their working timeslot. The appliance like electric kettle where the Length of Operation cannot accept pause and resume is uninterruptible. On requirement, appliance like washing machine where it's LoP can accommodate the break in its overall operation time. Thus the interruptible appliances play major role in reducing demand of electricity. Here, we consider the allocation of one timeslot completely rather considering the exact required appliance working 'LoP'. As the paper focuses on the peak load reduction of Demand Response program and the overall cost reduction of electricity consumers, only the possible load scheduling is expected from the consumer power consumption.

This work focuses on the optimal scheduling of resident appliances. The optimal appliances scheduling is formulated as an optimization problem and is solved by applying improved Genetic Algorithm (GA). In this improved GA entropy based selection, blended crossover and uniform mutation are performed. The simulation is performed for a single resident load profile with a set of appliances using Python programming. This work aims for reduction in the electricity cost to that of the Real time pricing. The constraints to be satisfied during this process are presented below:

1. Average power consumption per timeslot is represented in equation (9):

The appliances operation should meet the average power consumption that is specific to each timeslot of a day.

$$\sum_{\alpha=1}^{m} P_{\alpha}^{t} \leq avg\_power \tag{9}$$

2. Uncompromised consumption over 'n' timeslot is represented in equation (10):

During the period under consideration, the total power requirement of the consumer needs to be satisfied.

$$\sum_{t=1}^{n} \sum_{\alpha=1}^{m} P_{\alpha}^{t} == scheduled\_power \tag{10}$$

3. Possible delay start of reschedulable appliances is represented in equation (11):

If the appliance start time is within the range of peak time period then shift the appliance starts time after peak end time.

$$if\ (P_S \leq\ {ST_P}_j^t \leq\ P_E)\ then\ {ST_P}_j^t ==\ \ P_{E+1} \quad (11)$$

4. Run pre_started Reschedulable appliances is represented in Figure 12:

If an uninterruptable appliance has already started its working before the beginning of the peak time then it is allowed to continue its operation without any interrupt.

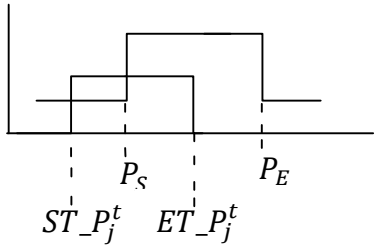

Figure 12. Run Pre_started Appliance

5. Possible delay start /advance start / break for interruptible appliances for raised peak is represented in equation (12):

If the operation of the appliances is within the range of peak timeslot, then based on the LoP either window is shifted left to or to the right.

$$LoP\_P_i^t\ =\ ST\_P_i^t\ \ to\ P_{S-1}\ \ and\ \ \ P_{E+1} to\ \ ET\_P_i^t \quad (12)$$

Where $P_S >\ {ST_P}_i^t$

The initial population is generated with certain number of individuals, where each individual represents a chromosome. The fitness function is used to evaluate the performance of the population. The selections are performed and apply the genetic parameters such as crossover and mutation. In every generation, the fitness of the individuals is evaluated. The parent chromosomes are selected for generating the off springs. The usual selection is based on the fitness proportion implemented using roulette wheel, Tournament, Rank, Stochastic Universal Sampling and Random selection. In this work, the selecting process is improvised by selection the population using Information Gain. Other methods like tournament selection, roulette wheel selection are stochastic in nature. To employ the deterministic approach in selecting the individual, the entropy in accordance to the fitness value is considered. The chromosome with more probability is selected as parent chromosome. Then these parent's are applied with genetic operators (crossover and mutation). The result of crossover is then performed

with mutation operation. The choice of mutation index is stochastic in nature. We propose a way to select a gene value of the individual. The gene value of a chromosome which does not contribute for the fitness value is considered rather than mutating a random gene.

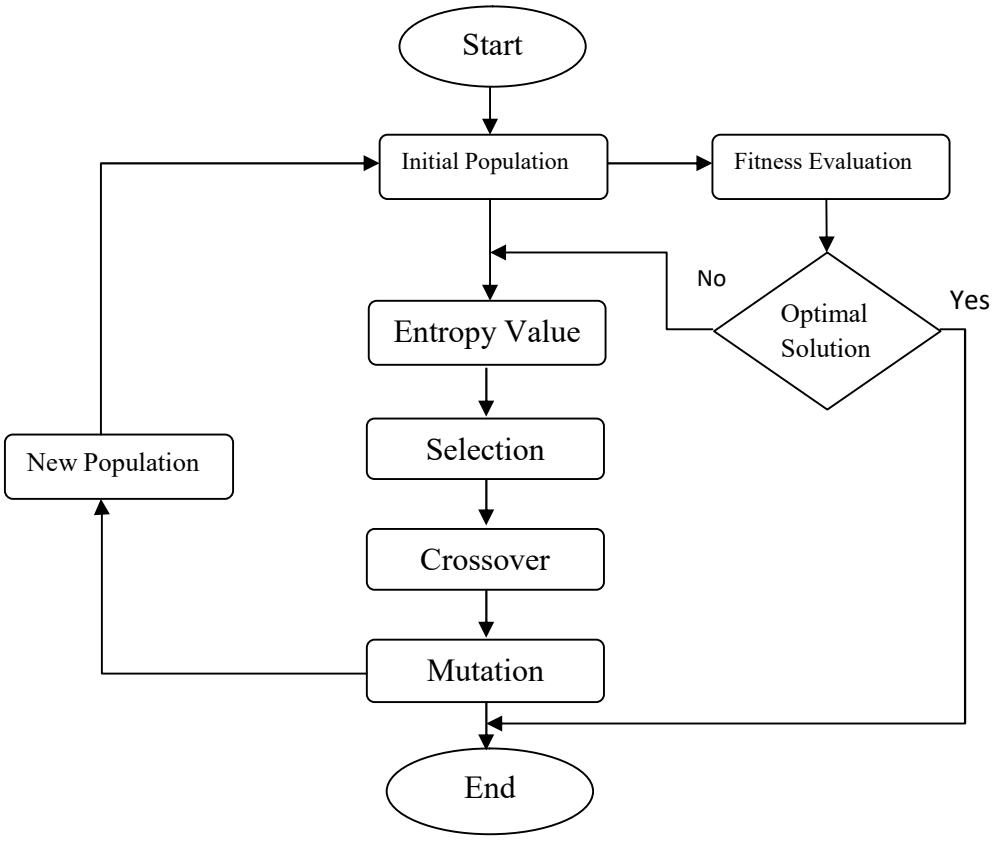

Figure 13. Improved GA flow diagram

The correlation coefficient value is used for selecting the mutation index. This permits to mutate the inactive gene values of the chromosome and to identify better individuals as population to participate in next generation. Figure 13 shows the overall flow of the proposed method. To bring in the entropy based approach in the selection process, the Information available in every chromosome is calculated with respect to the calculated fitness value. The fitness value is the objective function result generated for individual chromosome. Table 6 shows the parameter setting of Genetic algorithm. The individual are classified into two classes based on the fitness above and below the average fitness. The real coded GA is possible with Blended Crossover (BLX) method where the individuals are taking its natural form of representation as floating point values itself. Let X1 and X2 are the two parent chromosomes where X1 < X2. BLX create the children whose value is lying in the range {{X1 - α (X2 - X1)}……{X2 - α (X2 - X1)}}, where α is a constant to be decided so that children solution do not come out of the range

of domain of the said parameter. Two children C1 and C2 are generated using the equation (16) and (17). Another parameter 'γ' represented in the equation (18), has to be identified by utilizing the 'α' and a random number' r' in the range of (0.0, 1.0) both exclusive.

$$C1 = (1 - Y) * X1 + \gamma * X2 \quad (16)$$

$$C2 = (1 - Y) * X2 + \gamma * X1 \quad (17)$$

$$\gamma = (1 + 2\alpha)\, r - \alpha \quad (18)$$

To measure the association between the chromosome genes and the fitness, the Pearson correlation coefficient is used. It is the most widely used formula that calculates the 'r' value range from -1.0 to +1.0. It measures the linear correlation between the two variables. If the value is near +1.0, then it represents the positive correlation between the considered variables. If the value is near -1.0, then it represents the negative correlation between the variables. The correlation is considered as less, if the value '0'. One gene of the chromosome corresponds to one timeslot of the day. The correlation is calculated between the gene values and fitness value of the individual solution. The less correlated gene value index is identified as the mutant index. 10 interruptible appliances and 5 reschedulable appliances are considered for the experiment. The time of the day with 24 hours is considered as 96 timeslots where 15 minutes as one timeslot. The input variable is represented as two dimensional array. The test is performed to schedule the electric load for one day. The computation time is reduced the proposed method. From the simulation result, the time taken for performing the execution of standard GA takes 10.7 sec and the improved GA takes 7.54 sec. Thus it is shown in Figure 14.

Table 6. Genetic algorithm parameters

Population size	15
Number of iterations	30
Pc	0.7
Pm	0.01

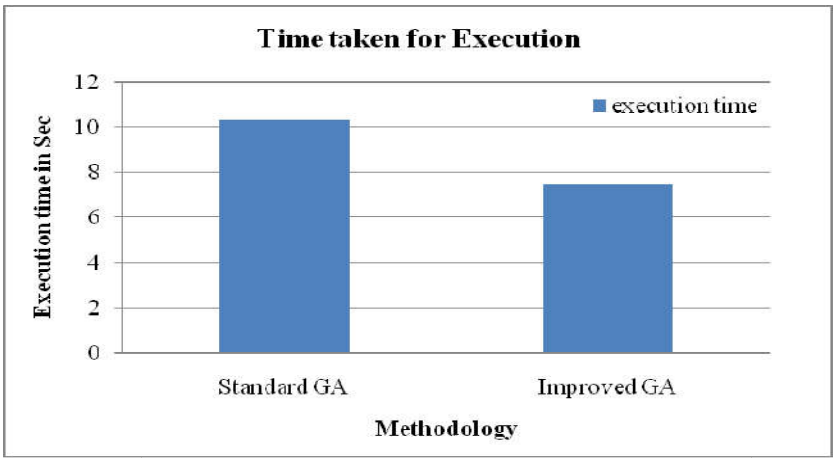

Figure 14. Algorithm Execution time

4. ORGANIZATION OF THESIS

Chapter 1 - *"Introduction"* presents content needed to understand the smart gird, AMI, Smart Meter, Application, and SMDAS. Machine learning model for building SMDAS, motivation of the research, objectives, contributions and organization of the thesis

Chapter 2 - *"Literature Survey"* reviews various prior research works in the field of Smart Meter Data Analytics for Demand Response program, Load scheduling and forecasting.

Chapter 3 - *"Load Prediction*

Estimating residential building energy consumption using demographic data describes the model of SMDAS to classify the consumers profile based on their consumption pattern and socio-economic data implemented using Deep Learning Neural Network. We perform the load prediction to plan the sustainable future electricity energy generation and to evaluate the model prediction accuracy. Also we provide the comparative results with conventional system models.

Chapter 4 - "*Theft Detection"*

Machine Learning Algorithm for Efficient Power Theft Detection approach describes the framework to cluster the consumers based on their consumption pattern and predict the future consumption using unsupervised and supervised learning algorithms. The bogus SM data's are generated and considered along with the real time genuine SM data to perform the evaluation of the developed framework. An Ensemble Machine Learning Model for Electricity Theft Detection approach describes a model that performs feature selection by finding the correlation among the

socio-economic data SM data of the consumers. Then an ensemble model is build using max voting technique on two trees based learning (Random Forest and XGradient Boosting) that predicts the consumer's consumption. The bogus SM data's are generated and considered along with the real time genuine SM data to perform the evaluation of the ensemble. Also the performance of the ensemble is evaluated against the regression models namely Multi - Linear Regression Model, Stepwise LR Model, Bayesian LR Model, K-Nearest Neighbor Model, Elastic-Net Model, Lasso and Elastic-Net Regularized Generalized Linear Model.

Chapter 5 -*"Demand Response"*

Load Profile Generation for Demand Response Program describes the framework for synthetic load profile generation using mathematical model to investigate the residential load consumption pattern and thereby evaluate the generated profile participation in Demand Response Program. Genetic Algorithm based Resident Load Scheduling for Electricity Cost Reduction performs cost minimization through load scheduling for residential appliances using Genetic algorithm. To optimizing the load schedule, 5 numbers of constraints are used to reduce the penalty cost of energy usage during demand time. Improved Genetic Algorithm for Optimal Demand Response in Smart Grid formulates an optimal load scheduling problem to make the consumer participate in the DR program and reduce their electricity cost. The optimized load schedule is proposed for resident using entropy based selection implementation in Genetic Algorithm. The performance evaluation is performed for the proposed model against the conventional GA.

Chapter 6 - "*Conclusion and Future work"* summarizes the research work to conclude the thesis and proposes the future scope as the extension of this research.

REFERENCES

1. A. Kavya., J. R. Matthew., B. Logan., and D. Deepak, Systematic Study of Data Requirements and AMI Capabilities for Smart Meter Analytics, *7th International Conference on Smart Energy Grid Engineering (SEGE), (2019*), 53–58.
2. Z. Aslam., N. Javaid., A. Ahmad., and SM. Gulfam, A Combined Deep Learning and Ensemble Learning Methodology to Avoid Electricity Theft in Smart Grids, *Energies,* **13**(21), (2020), 1–24.
3. B. Simona., B. Carlo Andrea., and M. Silvia, Socio-Economic Acceptability for Smart Grid Development – a Comprehensive Review, *Journal of Cleaner Production,* **131**, (2016), 399–409.

4. Y. Chen., H. Hou., Z. Zhang., L. Xianqiang., W. Jianjian., and T. Aihong., Review on Smart Meter Data Clustering and Demand Response Analytics, *12th IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC),* (2020), 1-6.
5. C. Maria Gabriel., and W. Fei, A Review of Non-Technical Loss Attack Models and Detection Methods in the Smart Grid, *Electric Power Systems Research* **199**, (2021), 107415 - 107421**.**
6. G. Hamid., and G. Reza, Smart Grid: The Electric Energy System of the Future, *Proceedings of the IEEE*, **99**(6), (2011), 917–21.
7. G. Ali Akbar., and G. Mohsen, Electrical Power and Energy Systems Detection of Illegal Consumers Using Pattern Classification Approach Combined with Levenberg-Marquardt Method in Smart Grid, *Electrical Power and Energy Systems,* **99,** (2018), 363–75.
8. H. Shih Che., L. Chan Nan., and L. Yuan Liang, Evaluation of AMI and SCADA Data Synergy for Distribution Feeder Modeling, *IEEE Transactions on Smart Grid,* **6(**4**)**, (2015), 1639–47.
9. D. P. Jenkins., S. Patidar., and S. A. Simpson, Synthesising Electrical Demand Profiles for UK Dwellings, *Energy and Buildings,* **76**, (2014), 605–14.
10. L. Xiufeng., G. Lukasz., G. Wojciech., and I. F. Ilyas, Benchmarking Smart Meter Data Analytics, *18th International Conference on Extending Database Technology, Proceedings* (*EDBT*), (2015), 385–96.
11. L. Renzhi., and H. Seung Ho, Incentive-Based Demand Response for Smart Grid with Reinforcement Learning and Deep Neural Network, *Applied Energy,* **236,** (2018), 937–49.
12. M. Martinez-Pabon., T. Eveleigh., and T. Bereket, Smart Meter Data Analytics for Optimal Customer Selection in Demand Response Programs, *Energy Procedia,* **107**, (2016), 49–59.
13. M. Fan Lin., and Z. Xiao Jun, A Profit Maximization Approach to Demand Response Management with Customers Behavior Learning in Smart Grid, *IEEE Transactions on Smart Grid*, **7**(3), (2016), 1516–1529.
14. N. Z. Sibo., Z. Ming., and L. Gengyin, Optimal Residential Community Demand Response Scheduling in Smart Grid, *Applied Energy,* **210**, (2018), 1280–1289.
15. S. Nimbargi., S. Mhaisne., S. Nangare., and M. Sinha, Review on AMI Technology for Smart Meter, *2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology*(ICAECCT), (2016), 21–27.
16. G. P. Nikolaos., E. Ozan., G. Anastasios., and J. P. S. Catalao, Optimal Household Appliances Scheduling under Day-Ahead Pricing and Load-Shaping Demand Response Strategies, *IEEE Transactions on Industrial Informatics,* **11**(6), (2015), 1509–19.

17. Q. Zhijian., L. Hanxin., W. Zixiao., X. Juan., P. Zhang., and H. Zeng, A Combined Genetic Optimization with AdaBoost Ensemble Model for Anomaly Detection in Buildings Electricity Consumption, *Energy and Buildings,* **248**, (2021), 111193 - 111199.
18. M. S. Saeed., M. Wazir Mustafa., N. N. Hamadneh, N. A. Alshammari., U. U. Sheikh., T. Ahmed Jumani., B. A. Khalid., and Ilyas Khan, Detection of Non-Technical Losses in Power Utilities —A Comprehensive Systematic Review, *Energies*, **13**(18), (2020), 1–25.
19. S. Ahmed., H. Maher., M. Elbayoumi., and S. Abdelhady, A Cuckoo Load Scheduling Optimization Approach for Smart Energy Management, *Energy Reports,* **7**, (2021), 4705–4721.
20. S. O. Tehrani., M. H. Yaghmaee Moghaddam., and Mohsen Asadi, Decision Tree Based Electricity Theft Detection in Smart Grid, *Proceeding of 4th International Conference on Smart Cities, Internet of Things and Applications* (SCIoT), (2020), 46–51.
21. M. L. Tuballa., and M. L. Abundo, A Review of the Development of Smart Grid Technologies, *Renewable and Sustainable Energy Review,* **59**, (2016), 710–25.
22. B. Yuce., Y. Rezgui., and M. Mourshed, ANN-GA Smart Appliance Scheduling for Optimised Energy Management in the Domestic Sector, *Energy and Buildings*, **111**, (2016), 311–25.

LIST OF PUBLICATION

1) **Jeyaranjani J** and Devaraj D**,** Machine learning algorithm for efficient power theft detection using smart meter data, *International Journal of Engineering and Technology (UAE)*,**7**,(2018), 900–904. **(Scopus Indexed) [Published]**

2) **Jeyaranjani J** and Devaraj D, Genetic Algorithm Based Resident Load Scheduling for Electricity Cost Reduction, *Lecture Notes in Electrical Engineering*, Springer, Singapore, **796**, (2021), 565–574. **(Scopus Indexed) [Published]**

3) **Jeyaranjani J** and Devaraj D, Load Profile Generation for DR Program, 2019 IEEE *International Conference on Intelligent Techniques in Control, Optimization and Signal Processing* INCOS, (2019), 1-5. **(Scopus Indexed) [Published]**

4) **Jeyaranjani J** and Devaraj D, Deep Learning based Smart Meter Data Analytics for Electricity Load Prediction, *2019 IEEE International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development (*INCCES), (2019), 1-5. **(Scopus Indexed) [Published]**

5) **Jeyaranjani J** and Devaraj D, Improved Genetic Algorithm for Optimal Demand Response in Smart Grid, *Journal of Sustainable computing, informatics and system* **35**, (2022), 1-8. (**SCI Impact Factor: 4.028) [Published]**

6) **Jeyaranjani J** and Devaraj D, An Ensemble Machine Learning Model for Electricity Theft Detection using Smart Meter Data, *IETE Journal of Research*. **(Scopus Indexed) [Accepted]**